# Towards Diagnosis of Rotator Cuff Tears in 3-D MRI Using 3-D Convolutional Neural Networks

**Mijung Kim** [1 2] **Ho-min Park** [1 2] **Jae Yoon Kim** [3] **Sofie Van Hoecke** [2] **Wesley De Neve** [1 2]

## Abstract

Torn rotator cuffs are the major cause of musculoskeletal pain in shoulders. These injuries, which are prevalent among overhead athletes and middle-aged adults and older individuals, may significantly degrade quality of life. A radiologist typically diagnose Rotator Cuff Tears (RCTs) by examining Magnetic Resonance Images (MRI) of a painful shoulder. Although deep learning algorithms are increasingly being used for analyzing medical images, they have not been deployed yet for the identification of RCTs, so to help them with making faster and more accurate diagnoses and medical decisions. Given this observation, we have developed a novel approach towards the diagnosis of RCTs in 3-D MRI, treating the diagnosis of RCTs as a 3-class classification problem (normal, partial-thickness tear, and full-thickness tear) that can be solved by leveraging 3-D Convolutional Neural Networks (CNNs). The proposed approach is able to achieve a diagnosis accuracy of 0.87 and an AUC score of 0.96, outperforming a baseline that makes use of wavelet-based features and gradient boosted decision trees, with this baseline reaching a diagnosis accuracy of 0.74 and an AUC score of 0.80.

## 1. Introduction

Rotator Cuff Tears (RCTs) are the major cause of musculoskeletal pain in shoulders. These injuries, as shown in Figure 1, mostly occur among middle-aged and older individuals, decreasing their quality of life by limiting shoulder movement with pain. Even young people may suffer from RCTs when playing overhead sports like baseball or tennis. Magnetic Resonance Imaging (MRI) of the shoulder is the *de facto* golden standard for the diagnosing shoulder injuries. However, shoulder MRI interpretation is a time-consuming and error-prone task. Although Computer-Assisted Diagnosis (CAD) tools have been developed for detecting brain tumors, breast cancer, and cardiac diseases, with many of these tools making use of conventional machine learning or deep learning, only a limited amount of work has thus far been done on CAD of RCTs (Litjens et al., 2017).

During the last decade, deep learning algorithms have outperformed conventional machine learning algorithms on most image segmentation and classification tasks. In particular, the application of Convolutional Neural Networks (CNNs) in medical imaging, from brain MRI to knee and eye fundus images, has shown promising results in terms of both lesion segmentation and disease diagnosis (Litjens et al., 2017; Usman & Rajpoot, 2017; Farahani et al., 2013; Gurusamy & Subramaniam, 2017; Abdi & Williams, 2010; Ashinsky et al., 2017).

Inspired by the above observations, we introduce a novel approach towards diagnosing RCTs in 3-D MRI using a deep learning algorithm. The major questions we set out to answer in our research are as follows:

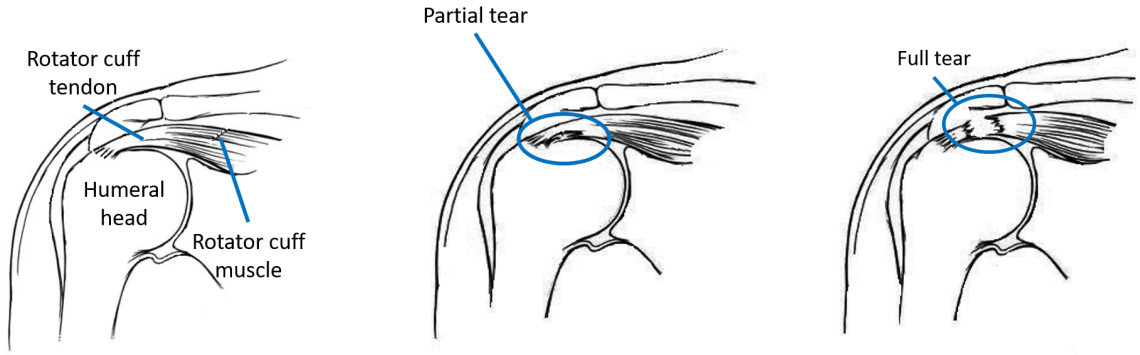
- How to leverage the information available along all dimensions in a 3-D MRI dataset?
- How to overcome the shortage of 3-D MRI scans for data-hungry deep learning algorithms?
- Can a CNN-based approach overcome data imbalance?

Our major contributions, which address the above-mentioned questions, are listed below:

- We present a preliminary research effort on diagnosing RCTs in 3-D MRI using 3-D CNNs, targeting three pathologies: normal, partial-thickness tear, and full-thickness tear.
- The proposed approach obtains a diagnosis accuracy of 0.87 and an AUC score of 0.96, outperforming several baseline machine learning approaches.

[1]Center for Biotech Data Science, Ghent University Global Campus, Korea [2]IDLab, ELIS, Ghent University, Belgium [3]Orthopedic Department, Chung-Ang University Hospital, Korea. Correspondence to: Mijung Kim <mijung.kim@ugent.be>.

*Workshop on Computational Biology at the International Conference on Machine Learning*, Long Beach, CA, 2019. 




(a) Normal tendon (b) Partial tear tendon (c) Full tear tendon

*Figure 1.* Anatomical comparison between different types of tendon.

## 2. Method

### 2.1. Dataset

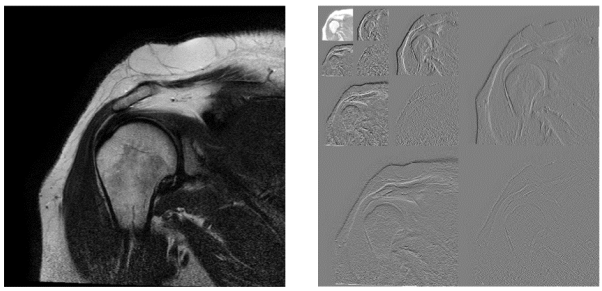
Shoulder MRI examinations were performed at Chung-Ang University Hospital between March 2010 and October 2018 (mean age 56 years; 941 [51%] female patients). These MRI examinations were then manually reviewed and labeled to curate a dataset. MRI scans were obtained using a 3.0 Tesla Achieva system built by Phillips, from March 2010 to June 2017. From June 2017 until October 2018, the MRI scans were obtained by a 3.0 Tesla Skyra system produced by Siemens AG Healthcare.

Considering tear severity, the dataset, which contains a total of 2,447 examinations, is divided into three different groups: 1,628 normal cases (66.5%), 157 cases of partial-thickness tears (6.4%), and 662 cases of full-thickness tears (27.1%). By proportion, the training set, the validation set, and the testing set are approximately 8 (1,963 examinations):1 (242 examinations):1 (242 examinations). All sets have the same proportion of normal cases, partial-thickness tear cases, and full-thickness tear cases. More detailed statistics can be found in Table 1.

### 2.2. Model

To extract images, we first pre-processed the raw MRI data. Next, to extract features, we used two different methods: a wavelet-based approach and an approach based on a 3-D CNN. Finally, we used the extracted features to classify the input into three different pathologies.

**Preprocessing:** From the original MRI Digital Imaging and Communications in Medicine (DICOM) files, we extracted Portable Network Graphics (PNG) images using Python and the pydicom library. Next, to make sure that the number of slices per examination is the same, we selected 16 slices from the center of each shoulder MRI examination. Indeed, a single DICOM file usually contains 200 to 500 slices, given that an MRI examination is typically performed using different image types (e.g., T1-weighted images, T2-weighted images) and different directions (coronal, axial, and sagittal), so to facilitate a diagnosis that is as accurate as possible. As such, based on the advice of orthopedic surgeons, we extracted the middle 16 slices of the T2-weighted coronal slices, as these are known to be the most important for detecting RCTs.



(a) Original (b) Wavelet transformed

*Figure 2.* Comparison between an MRI image and its wavelet transformed version. To extract features, three vertical and horizontal filter banks were applied to each MRI image. After dimension reduction using PCA, the obtained features were used for the purpose of classification.

**Wavelet transform:** To perform a comparative investigation of the effectiveness of the proposed model, we used the feature extraction method of (Nayak et al., 2016). This method extracts a 13-D feature vector per image by applying a three-level wavelet decomposition and PPCA (Probabilistic PCA) to brain MRI images, subsequently applying an AdaBoost classification model to detect brain lesions. Given the aforementioned approach, and as shown in Figure 2, we extracted features by applying the same technique to our

*Table 1.* Summarizing statistics for the shoulder MRI datasets used for training, validation, and testing. The proportion of each class is approximately the same in each dataset.

| Statistics | Training | Validation | Testing |
|---|---|---|---|
| Total number of examinations (%) | 1,963 (100) | 242 (100) | 242 (100) |
| - Number of normal examinations (%) | 1,308 (66.6) | 160 (66.1) | 160 (66.1) |
| - Number of partial-thickness tear examinations (%) | 125 (6.4) | 16 (6.6) | 16 (6.6) |
| - Number of full-thickness tear examinations (%) | 530 (27.0) | 66 (27.3) | 66 (27.3) |
| Total number of patients | 1,847 | 231 | 228 |
| - Number of female patients (%) | 942 (51) | 115 (49) | 134 (58) |
| - Age mean by the number of patients ($\pm std.$) | 56 ($\pm$14.8) | 57 ($\pm$14.9) | 56 ($\pm$14.6) |

MRI scans. However, unlike the model of (Nayak et al., 2016), we need to distinguish between three pathologies using 16 MRI slices rather than a single MRI slice. Therefore, we obtained a 13-D feature vector for each of the 16 slices extracted per patient, subsequently concatenating these feature vectors. This resulted in the creation of a 208-D feature vector for each of the $2,447$ MRI examinations, originally having a size of $16 \times 255 \times 255 \times 2,447$. Next, we used these feature vectors to train various kinds of machine learning models. For parameter tuning, we used a grid search approach.

**3-D CNN:** As briefly mentioned in the introductory section, a CNN is currently a popular choice for building CAD tools. This is mainly because a CNN convolves filters over the given input image, hereby capturing spatial relations between pixels. By changing the size of the filters and the number of layers, a CNN is able to improve its effectiveness in terms of image classification accuracy. However, normal 2-D convolutional filters are not able to take advantage of the temporal information available in 3-D input samples such as video clips. To overcome this issue in the context of video action recognition, Facebook Research for instance introduced a 3-D CNN approach (Tran et al., 2015).

Considering the 3-D characteristics of the MRI scans available, we also adopted a 3-D CNN approach, using as base architecture the 3-D CNN model introduced by Facebook Research for the purpose of video action recognition, consisting of eight 3-D convolutional layers. However, given that our shoulder MRI dataset does not have enough examinations to train the aforementioned 3-D CNN model from scratch, we adopted transfer learning to mitigate the risk of overfitting and to reduce the time needed for training (Torrey & Shavlik, 2010). First, to create a pre-trained model, we trained the 3-D CNN model from scratch on the UCF101 video dataset (Soomro et al., 2012), coming with 101 different classes in support of video action recognition. During this phase, we used the data augmentation methods explained below, as well as the hyperparameter values that can be found in Table 2. The model with the lowest validation loss was selected for fine-tuning on our shoulder MRI dataset.

Given the pre-trained model created during the above-discussed phase, we trained this model again on our shoulder MRI dataset, this time using a different set of hyperparameter values. These values can also be found in Table 2. During training, each slice of a clip was resized to $112 \times 112$ pixels, with each pixel having 3 color components. The obtained clip was then randomly augmented using several methods - rotation, horizontal flipping, shifting, Gaussian filter and noises. As for the rotation, the rotation angle is randomly selected between -15 and 15 degrees, flipped horizontally with a 50% probability, shifted randomly between -10 and 10 pixels, the subject of Gaussian filters (blurring) with a random window size ($1 \times 1, 3 \times 3, 5 \times 5, 7 \times 7$, and $9 \times 9$), and the subject of Gaussian noise ($\mu$:0, $\sigma^2$:1). Gamma adjustment used random gamma between 0.5 and 1.0. All training examples were augmented using the same random parameters per clip. Figure 4 provides a more detailed overview of the different types of data augmentation applied. Each of the obtained training examples was then fed to the pre-trained network model, making it possible to fine-tune the weights of this model. The softmax function was used to predict a particular class. Optimization of the model was done using the categorical cross-entropy loss. The optimizer used was Adaptive Moment Estimation (Adam) (Kingma & Ba, 2014).

**Implementation:** We implemented our 3-D CNN model using Python 2.7 with TensorFlow r1.12 and OpenCV 3.2 package for data augmentation, running this model using two Intel(R) Xeon(R) E5-2620 2.4GHz CPUs and an NVIDIA GeForce GTX TITAN X GPU.

## 3. Experimental Results

As shown in Table 3, the overall diagnosis accuracy of our approach is 0.87, which is significantly higher than any of the other machine learning approaches implemented. For example, the second highest diagnosis accuracy was obtained by a model using gradient boosted decision trees, obtaining a value of 0.74. The micro-averaged AUC score of our ap-

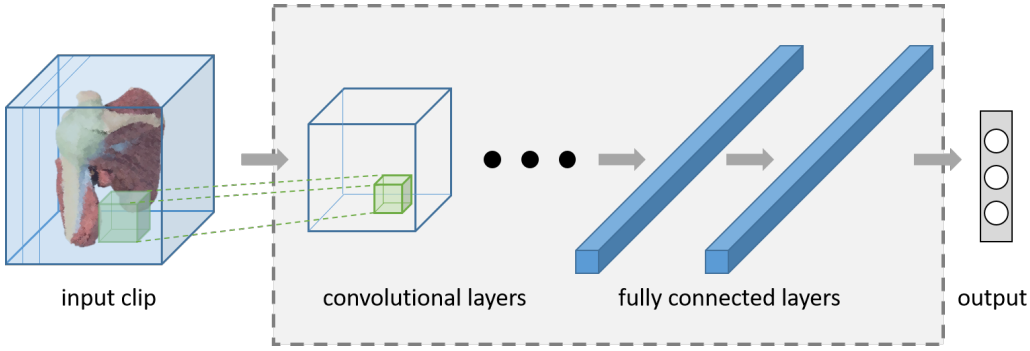



*Figure 3.* Augmented input clips of 16 images are used for fine-tuning a 3-D CNN that consists of eight 3-D convolutional layers. The layers within the dashed bounding box were pre-trained using the UCF101 video dataset. The obtained weights were then transferred and fine-tuned using our shoulder 3-D MRI dataset. The resulting 3-D CNN is then able to determine whether or not an unseen input clip contains torn tendon slices.

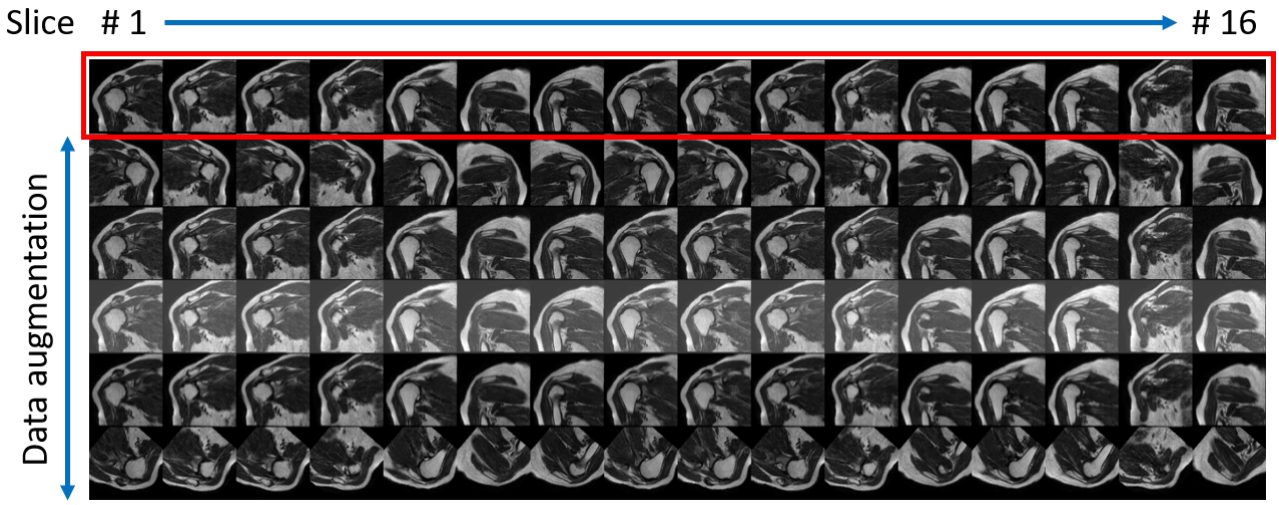



*Figure 4.* An example of data augmentation applied to an input clip of 16 MRI images. The first row in the red rectangle shows the original input images, from the left to the right in order. Starting from the second row, we can see the employed data augmentation techniques: horizontal flipping, Gaussian noise, gamma adjustment, Gaussian blur, and rotation. Each augmentation technique is activated with 0.5 probability. During training, we randomly selected from zero to all the methods and applied to the input clip.

*Table 2.* Hyperparameter settings for the UCF101 training set and the shoulder MRI dataset.

| Hyperparameters | Dataset | |
|---|---|---|
| | UCF101 | Shoulder MRI |
| Learning rate | 1e-3 | 1e-5 |
| Batch size | 16 | 16 |
| Clip size | 16 | 16 |
| Epoch (early stopping) | 49 | 25 |

proach was also the highest, coming with a value of 0.96. The micro-averaged AUC score of the other machine learning algorithms varied from 0.64 to 0.80. Furthermore, our 3-D CNN model outperformed the other machine learning approaches in terms of precision, recall, and F1 score.

We also evaluated our model to see how it deals with an imbalanced dataset. First, we plotted the ROC-AUC curve of each class in Figure 5. Compared to the normal and full-thickness tear classes, which have an AUC score of 0.93 and 0.96, respectively, the partial-thickness tear class comes with an AUC score of 0.69. To enable more in-depth analysis, we adopted two other measures, namely the Precision-Recall curve and the confusion matrix. First, compared to an ROC-AUC curve, a Precision-Recall curve is a more useful measure of prediction effectiveness when the classes of the given dataset are significantly imbalanced, as is the case for our dataset (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). Among all classifiers implemented, we made Precision-Recall plots for the model using a random forest and for the model using gradient boosted decision

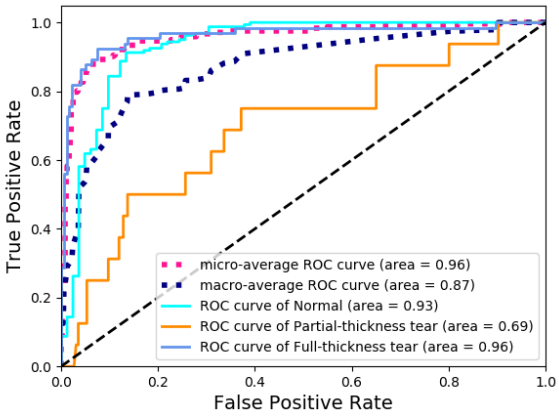



*Figure 5.* ROC-AUC curve per class.

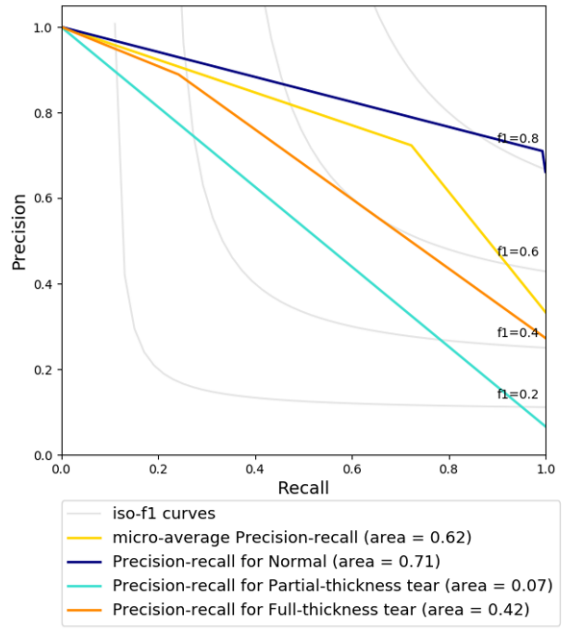



*Figure 6.* Precision-recall curve for each class, as obtained for the random forest model.

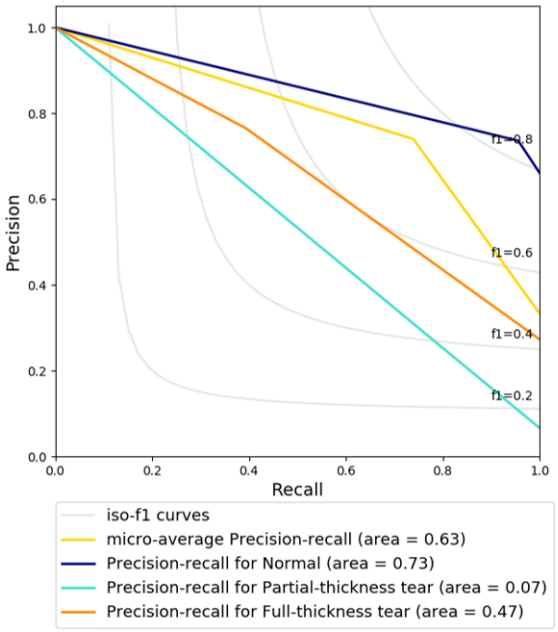



*Figure 7.* Precision-recall curve for each class, as obtained for the model using gradient boosted decision trees.

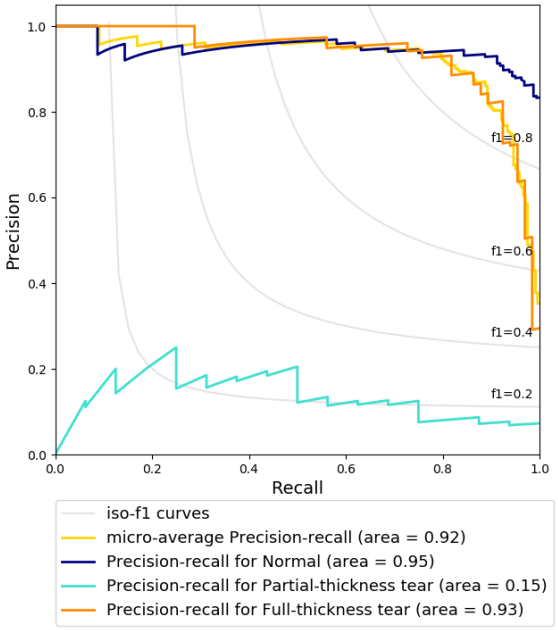



*Figure 8.* Precision-recall curve for each class, as obtained for our 3-D CNN.

trees, given that both models demonstrated to have similar levels of effectiveness in Table 3. Compared to both machine learning classifiers in Figure 6 and Figure 7, the 3-D CNN in Figure 8 shows a significantly higher Precision-Recall score for the normal and full-thickness tear classes. However, for the partial-thickness tear class, which encompasses about 6.6% of all test examples, the Precision-Recall score was close to zero for all models.

Furthermore, the confusion matrix of the 3-D CNN in Figure 9 shows that our proposed model recognizes well the difference between the normal class and the full-thickness tear class. However, 88% of the partial-thickness tear samples are misclassified as normal and 12% of the partial-thickness tear samples are misclassified as full-thickness tear. Overcoming this data imbalance problem remains as a future work item.

*Table 3.* Overall comparison of the effectiveness of the different classifiers used. Each of the values is averaged based on the total number of test examples. M-AUC denotes the macro-averaged AUC score and m-AUC denotes the micro-averaged AUC score.

| Model | Accuracy | Precision | Recall | F1 score | M-AUC | m-AUC |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.72 | 0.67 | 0.72 | 0.67 | 0.59 | 0.79 |
| AdaBoost | 0.66 | 0.44 | 0.66 | 0.53 | 0.50 | 0.75 |
| K-Nearest Neighbors | 0.67 | 0.61 | 0.67 | 0.63 | 0.56 | 0.76 |
| Decision Tree | 0.68 | 0.63 | 0.68 | 0.65 | 0.60 | 0.76 |
| Random Forest | 0.73 | 0.72 | 0.73 | 0.66 | 0.58 | 0.80 |
| Multi-layer Perceptrons | 0.71 | 0.66 | 0.71 | 0.66 | 0.58 | 0.79 |
| Gaussian NB | 0.52 | 0.67 | 0.52 | 0.56 | 0.61 | 0.64 |
| Quadratic Discriminant Analysis | 0.57 | 0.55 | 0.57 | 0.56 | 0.54 | 0.68 |
| Gaussian Process | 0.61 | 0.60 | 0.61 | 0.60 | 0.57 | 0.71 |
| XGBoost | 0.74 | 0.69 | 0.74 | 0.69 | 0.61 | 0.80 |
| **Our approach** | **0.87** | **0.81** | **0.87** | **0.84** | **0.87** | **0.96** |

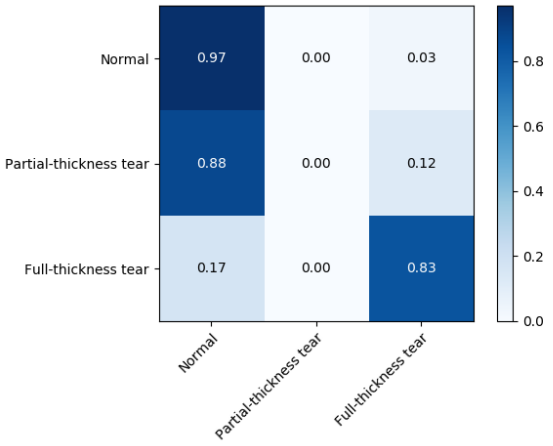



*Figure 9.* Confusion matrix for each class. The Y-axis shows the true labels and the X-axis shows the predicted labels.

## 4. Conclusions and Future Work

In this paper, we presented and evaluated a 3-D CNN model for diagnosing RCTs in 3-D shoulder MRI, treating the problem of RCT diagnosis as a three-class classification problem. Preliminary experimental results demonstrated that our model is able to achieve an overall diagnosis accuracy of 0.87 and an AUC score of 0.96. However, the proposed approach might be biased towards the dataset collected at Chung-Ang University Hospital, requiring validation using an external dataset. Moreover, the 3-D CNN model used, together with all other machine learning models implemented, needs to be improved so to be able to better deal with an imbalanced dataset.

In future research, we plan to perform a more extensive exploration of deep learning models using different 3-D CNN approaches, with the goal of improving the diagnosis effectiveness. Furthermore, we need to examine the generalizability of the predictive models used by testing on an external dataset. Finally, by overcoming the imbalanced dataset problem, we should be able to move from coarse- to fine-grained RCT classification.

## Acknowledgements

The research efforts described in this paper were funded by Ghent University, Ghent University Global Campus, Flanders Innovation & Entrepreneurship (VLAIO), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

## References

Abdi, H. and Williams, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Ashinsky, B. G., Bouhrara, M., Coletta, C. E., Lehallier, B., Urish, K. L., Lin, P.-C., Goldberg, I. G., and Spencer, R. G. Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative. *Journal of Orthopaedic Research*, 35(10):2243–2250, 2017.

Davis, J. and Goadrich, M. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240. ACM, 2006.

Farahani, F., Reyes, M., Menze, B., Gerstner, E., Kirby, J., and Kalpathy-Cramer, J. NCI-MICCAI 2013 challenge on multimodal brain tumor segmentation (BRaTS). *The challenge database contains fully anonymized images from the following institutions: ETH Zurich, University of Bern, University of Debrecen, and University of Utah*

*and publicly available images from the Cancer Imaging Archive (TCIA)*, 8, 2013.

Gurusamy, R. and Subramaniam, V. A machine learning approach for MRI brain tumor classification. *Computers, Materials & Continua*, 53(2):91–108, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

Nayak, D. R., Dash, R., and Majhi, B. Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests. *Neurocomputing*, 177:188–197, 2016.

Saito, T. and Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10 (3):e0118432, 2015.

Soomro, K., Zamir, A. R., and Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Torrey, L. and Shavlik, J. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pp. 242–264. IGI Global, 2010.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

Usman, K. and Rajpoot, K. Brain tumor classification from multi-modality MRI using wavelets and machine learning. *Pattern Analysis and Applications*, 20(3):871–881, 2017.